

Guía 4: Hashing

1. Al construir una función de hashing perfecta y mínima con el método visto en clase explique por que es necesario que M sea mayor a N .
2. Encontrar el m mínimo y los parámetros a y b de forma tal que la función de hashing $ax + b \bmod m$ sea perfecta para las siguientes claves: 1,3,5,12
3. Tenemos vectores en 4 dimensiones y usamos "the hashing trick" usando el método de una única función de hashing (es decir sin signo) para reducirlos a 3 dimensiones. Sabemos que la matriz asociada a la función de hashing usada es la siguiente (por filas): $[1,0,0; 0,0,1; 0,1,0; 1,0,0]$. Se pide construir la función de hashing $h(x)$ equivalente a la matriz presentada.
4. Dada la siguiente función de hashing que pertenece a la familia Universal de Carter-Wegman para números enteros: $h(x) = [(4*x + 3) \bmod 13] \bmod 5$. Usamos h para construir un esquema FKS para las siguientes claves: 20,40,70,10,100. Indicar la estructura final resultante y la en caso de ser necesario la segunda función de hashing a usar para el segundo nivel teniendo en cuenta que debe ser pertenecer a la familia $[(a*x + b) \bmod 13] \bmod m$
5. Tenemos un total de 10.000 claves de 32 bytes c/u. Si usamos el esquema FKS y la primer tabla tiene 1000 posiciones. ¿Cuánto espacio necesitamos en total para almacenar las 10.000 claves?
6. Supongamos que asignamos a cada letra del alfabeto un número de la forma $A=1, B=2, C=3, \dots$ Proponemos como función de hashing sumar el valor correspondiente a cada carácter del string y luego tomar el módulo con un cierto número primo p . Analizar la función propuesta indicando:
 - a) Cantidad de colisiones
 - b) Facilidad de encontrar sinónimos
 - c) Eficiencia
 - d) Efecto avalancha
7. Determinar si las siguientes afirmaciones son V / F justificando la respuesta:
 - a. Una función de hashing criptográfica produce muy pocas colisiones. \checkmark
 - b. La construcción de DavisMeyer es necesaria para que la función de hashing produzca muy pocas colisiones. F

↓ DIFÍCIL HALLAR X?

UN CAMBIO MIN EN --



- c. El efecto avalancha se produce cuando muchas claves hashean a un mismo valor generando muchas colisiones .
- d. Una función de hashing para strings debe poder generar el resultado muy rápidamente.
- e. Utilizando una función de hashing con resolución de colisiones Hopscotch de distancia máxima 4 me aseguro insertar como mínimo 4 sinónimos.
- f. Dado que una función de hashing criptográfica tiene pocas colisiones es ideal para asegurarnos buscar claves rápidamente y su uso es recomendable en la mayoría de los casos.
- g. Si las claves a hashear son numéricas y aleatorias entonces no es necesario que el número m sea primo en la función de hashing . $a^*x \bmod m$.
- h. Aumentar la cantidad de funciones de hashing o la cantidad de registros por bucket en el método del cuckoo sirve para reducir la cantidad de accesos promedio que necesitamos para recuperar una clave.
- i. Si H es una familia de funciones de hashing universal definida por $h(x,a)$ en donde "a" es un parámetro entonces la familia H_2 definida $h_2(x,a) = h(x,a) \bmod p$ también es universal.
- j. Una función de hashing criptográfica como SHA-256 genera menos colisiones que una función genérica como Jenkins pero es mucho mas lenta.
- k. El tiempo necesario para generar una función de hashing perfecta usando Hash & Displace depende de la cantidad de colisiones que genere la primer función de hashing.
- l. Dado un conjunto de claves numéricas de 32 bits la función de hashing $x \% 1000$ es igual de buena que la función de hashing $x \% 1001$.
- m. Una función de hashing puede ser perfecta y sin embargo no servir como función de hashing criptográfica.

GUÍA 4 - HASHING

(1)

UNA FUNCION DE HASHING ES PERFECTA Y MINIMA CUANDO ES $O(1)$ PARA CONSULTAS Y $O(M)$ ESPACIO

(2)

HASH PERFECTO → NO HAY COLISIONES

$$H(x) = Ax + B \bmod M \rightarrow \text{UN BUCKET PARA CADA ELEMENTO}$$

$M \rightarrow$ CANTIDAD DE BUCKETS → TENGO 4 ELEMENTOS

$$\text{ELEMENTOS} = \{1, 3, 5, 12\}$$

$$\downarrow \text{MIN} = 4$$

$$0 = A \cdot 1 + B \bmod 4$$

NO HAY NINGÚN VALOR

$$1 = A \cdot 3 + B \bmod 4$$

CON $M=4$ QUE CUMPLA

$$2 = A \cdot 5 + B \bmod 4$$

\downarrow
ME FIJO CON $M=5$

$$3 = A \cdot 12 + B \bmod 4$$

Y CUMPLE CON

$$A=1, B=0$$

$H(x) = X \bmod 5$ ES UN HASH PERFECTO
PARA ESTE CONJUNTO

3

VECTORES 4 DIM $V = [W, X, Y, Z]$
 $\begin{matrix} 0 & 1 & 2 & 3 \end{matrix}$

"THE HASHING TRICK" QUIERO REDUCIRLO A 3 DIM

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

QUEDEMOS LA $H(X)$ QUE
REPRESENTA ESTA MATRIZ

LA MATRIZ HACE REFERENCIA
A LAS POSICIONES QUE ESTABAN

HAY 1 EN POS(0) LOS ELEMENTOS

POS (2)	W+Z	Y	X
POS (1)			
POS (0)	0	1	2

HASHEO LOS INDICES!!

ENTONCES YO QUIERO • $H(0) = 0$

• $H(1) = 2$

$H(3) = 0$ • $H(2) = 1$

$$H(x) = Ax + B \pmod{P} \pmod{M}$$

$M = 3$ PORQUE QUIERO REDUCIR LA DIM EN 1

$P \rightarrow 3$, $\rightarrow P = 3$ CUANDO SON = SACO UN MOD

$$H(x) = Ax + B \pmod{3}$$

$$0 = B \pmod{3} \quad B = 0 \quad A = 2 \rightarrow \text{CUMPLE!}$$

$$0 = 3A + B \pmod{3}$$

$$H(x) = 2x \pmod{3}$$

$$2 = A + B \pmod{3}$$

$$1 = 2A + B \pmod{3}$$

$$\text{MOD}(A, B) \quad A < B \rightarrow \text{MOD} = A$$

NOTA

$$A > B \quad (A : B) \rightarrow \text{LO ENTENDO} \rightarrow A - z$$

$$\times B = z$$

4

ELEMENTOS = { 20, 40, 70, 10, 100 }

APLICAR FUNCION DADA H(X).

$$H(x) = ((4x + 3) \text{ MOD } 13) \text{ MOD } 5$$

• $H(20) = (4 \cdot 20 + 3 \text{ MOD } 13) \text{ MOD } 5$

$$H(20) = (83 \text{ MOD } 13) \text{ MOD } 5$$

$$H(20) = (5) \text{ MOD } 5$$

$$H(20) = 0$$

• $H(40) = (4 \cdot 40 + 3 \text{ MOD } 13) \text{ MOD } 5$

$$= (163 \text{ MOD } 13) \text{ MOD } 5$$

$$= (7) \text{ MOD } 5$$

$$= 2$$

• $H(70) = (4 \cdot 70 + 3 \text{ MOD } 13) \text{ MOD } 5$

$$= (283 \text{ MOD } 13) \text{ MOD } 5$$

$$= (10) \text{ MOD } 5$$

$$= 0$$

• $H(10) = (4 \cdot 10 + 3 \text{ MOD } 13) \text{ MOD } 5$

$$H(10) = (43 \text{ MOD } 13) \text{ MOD } 5$$

$$H(10) = (4) \text{ MOD } 5$$

$$H(10) = 4$$

• $H(100) = (100 \cdot 10 + 3 \text{ MOD } 13) \text{ MOD } 5$

$$= (2) \text{ MOD } 5$$

$$= 0$$

INDICE	CLAVE	HASH 2	SEGUNDA TABLA
0	20, 70, 100	$(x+2 \text{ MOD } 13) \text{ MOD } 9$	{ 20, 70, 100 }
1			
2	40	$(x \text{ MOD } 1)$	{ 40 }
3			
4	10	$x \text{ MOD } 1$	{ 10 }

EN LA POS 0 TENGO COLISION DE 3 VALORES $M_0^2 = 9$

$H_2(x)$ TIENE QUE SER DEL TIPO

$$H_2(x) = (Ax + B \text{ MOD } 13) \text{ MOD } M$$

$$H_2(x) = (x + 2 \text{ MOD } 13) \text{ MOD } 9$$

ESTO SERIA
MEJOR ELIENDO
 $M_i^2 \rightarrow k +$
CERCANO

$$\begin{aligned}
 H_2(20) &= (20 + 2 \text{ MOD } 13) \text{ MOD } 9 \\
 &= (22 \text{ MOD } 13) \text{ MOD } 9 \\
 &= 9 \text{ MOD } 9 \\
 &= 0 \checkmark
 \end{aligned}$$

$$\begin{aligned}
 H_2(70) &= (70 + 2 \text{ MOD } 13) \text{ MOD } 9 \\
 &= (72 \text{ MOD } 13) \text{ MOD } 9 \\
 &= 7 \text{ MOD } 9 \\
 &= 7 \checkmark
 \end{aligned}$$

$$\begin{aligned}
 H_2(100) &= (100 + 2 \text{ MOD } 13) \text{ MOD } 9 \\
 &= (102 \text{ MOD } 13) \text{ MOD } 9 \\
 &= 11 \text{ MOD } 9
 \end{aligned}$$

$$H_2(100) = 2 \checkmark$$

Guía 5: LSH

1. Tenemos la siguiente tabla representando el valor de 6 minhashes para tres documentos:

	D1	D2	D3
MH1	1	0	1
MH2	3	1	3
MH3	1	2	1
MH4	2	2	3
MH5	0	0	2
MH6	0	0	2

Se usa $b=2$ y $r=3$. Se decide usar 7 buckets para cada banda.

Encontrar una única función de hashing perteneciente a una familia universal LSH(r_1, r_2, r_3) de forma tal que en la primera banda solo D1 y D3 sean candidatos a ser similares pero en la segunda banda los tres documentos sean candidatos a ser similares.

2. Si se tiene la siguiente familia de funciones LSH (0.15,0.85,0.85,0.15) indique de qué forma quedaría amplificada usando $r=3$ y $b=4$. Finalmente interprete el resultado de la familia amplificada indicando qué cantidad de falsos positivos o falsos negativos se producirían.
- \checkmark AUMENTAR R Y B \rightarrow ACHICAR POSIBILIDAD DE COWISION
3. Usando LSH en una construcción de 5 ANDs y 3 ORs se observa que algunos pares de documentos que deberían ser candidatos a similares no lo son. ¿Qué cambiaría?
- FALSOS NEGATIVOS
4. Se quiere aplicar LSH a un conjunto de documentos para encontrar los pares de documentos más similares. Queremos que si $J(D1, D2) \geq 0.7$ entonces la probabilidad de que D1 y D2 sean candidatos sea ≥ 0.9 y queremos que si $J(D1, D2) \leq 0.5$ entonces la probabilidad de que sean candidatos sea ≤ 0.3 . Indique cuántas funciones minhash usaría y qué combinación de AND y OR usaría para lograr lo pedido.
5. Se quiere construir una función LSH usando Jaccard que detecte aquellos documentos cuya semejanza esté entre 0.8 y 1.0. Vamos a pedir que si dos documentos tienen semejanza 0.9 o mayor la probabilidad de detectarlos sea 0.95 y que si dos documentos tienen semejanza 0.8 o menor la probabilidad de detectarlos sea inferior a 0.2. Construir la función LSH pedida usando la menor cantidad de funciones de hashing posible, indicar r y b . Reflexione sobre lo que pasó en este ejercicio.
- ? 6. Se tienen los siguientes puntos en el plano: (2,3) (3,4) (24,30) (21,32). Sean el siguiente vector al azar: (5,3). Indique cuál debería el valor de w para que al aplicar LSH para la distancia euclídea los puntos 1 y 2 sean semejantes pero los puntos 3 y 4 no.

7. Usamos LSH para encontrar documentos parecidos a un documento consulta. Desafortunadamente nuestra función LSH no está encontrando varios documentos que son similares a los buscados y esto es importante para nuestro problema. Indique diferentes formas de solucionar este problema.
8. Usando la distancia de Jaccard y 36 minhashes se quiere comparar el efecto de usar 6 construcciones OR y luego 6 AND contra usar primero 6 construcciones AND y luego 6 OR. ¿En qué casos tendremos mas falsos positivos y en que casos mas falsos negativos? Si fijamos $d_1=0.2$ y $d_2=0.5$ ¿cuál es la probabilidad de que dos documentos sean candidatos en cada caso?
9. Usamos LSH para la distancia de Jaccard para comparar frases breves usando 4-shingles con 6 funciones de hashing que agrupamos en 3 construcciones OR de 2 construcciones AND cada una. Queremos obtener los strings que sean al menos 80% semejantes a "use the force". Describa detalladamente todos los pasos necesarios para encontrar las frases que cumplan con lo pedido.
10. Dados los vectores: $x=[1,3,-1,2]$; $y=[-1,-2,-1,-1]$, $z=[2,4,-1,3]$ y los hiperplanos aleatorios: $r_1 = [+1,-1,+1]$ $r_2=[+1,+1,-1]$ $r_3=[-1,-1,-1]$, $r_4=[+1,+1,+1]$. Queremos usar 3 (tres) hiperplanos para aproximar el coseno entre los vectores usando LSH. ¿Cuáles son los 3 hiperplanos que hay que elegir entre los 4 propuestos? Justifique adecuadamente
11. Sean los siguientes vectores en 5 dimensiones: $v_1 = [4 \ 4 \ -5 \ -2 \ 3]$; $v_2 = [-3 \ -2 \ -4 \ 5 \ 0]$; $v_3 = [3 \ 2 \ -1 \ -2 \ 1]$. Y sean los siguientes 6 hiperplanos aleatorios: $r_1 = [1 \ 1 \ 1 \ 1 \ -1]$; $r_2 = [-1 \ 1 \ -1 \ -1 \ -1]$; $r_3 = [1 \ -1 \ -1 \ -1 \ -1]$; $r_4 = [1 \ -1 \ -1 \ -1 \ 1]$; $r_5 = [1 \ -1 \ -1 \ -1 \ 1]$; $r_6 = [-1 \ 1 \ 1 \ -1 \ 1]$. Se pide comparar las alternativas $r=3$, $b=2$ vs $r=2$, $b=3$ indicando en cada caso que colisiones se producirían.
12. Si la probabilidad de que dos vectores colisionen usando un único hiperplano es mayor a 0.95.
- ¿Cuál es el ángulo máximo entre los vectores?
 - De un ejemplo de un hiperplano para el cual dos vectores que están a la distancia indicada en el punto anterior no colisionen.

$$1 - \frac{\theta}{180} = 0,95$$

DESPEJAS θ

VECTORES
QUE HAGA
QUE EL BIT
SEA \neq .

GUÍA 5 - LSH

(1)

	D1	D2	D3	B = 2 R = 3
MH1	1	0	1	
MH2	3	1	3	+ BUCKETS
MH3	1	2	1	M = 7

MH4	2	2	3	
MH5	0	0	2	
MH6	0	0	2	

CONDICIONES

$$\bullet D_1 = D_3$$

$$H(131) = H(131)$$

$$H(131) \neq H(012)$$

$$\bullet D_1 = D_2 = D_3$$

$$H(200) = H(322)$$

$$H \in \mathcal{H} \Rightarrow H(x) = \sum_{i=0}^{R-1} (A_i * x_i \text{ } (\text{MOD } P)) \text{ } (\text{MOD } M)$$

CLAVES - 1

COMO M YA ES UN NÚMERO PRIMO

$$H \in \mathcal{H} \Rightarrow H(x) = \sum_{i=0}^{R-1} (A_i * x_i) (\text{MOD } M)$$

$$H(X) = AX + BY + CZ \bmod 7$$

$$\bullet H(200) = H(322)$$

$$2A \bmod 7 = 3A + 2B + 2C \bmod 7$$

$$\bullet H(131) \neq H(012)$$

$$2A + 3B + C \bmod 7 \neq B + 2C \bmod 7$$

RESUELVO POR FUERZA BRUTA

$$A = 1, B = 2, C = 1 \quad (\text{HAY QUE PROBANDO}, \\ \text{ESTOS ME LOS DIJO TOMI})$$

$$H(X) = X + 2Y + Z \bmod 7$$

(2)

$$LSH(0.15, 0.85, 0.85, 0.15)$$

$$R = 3 \quad B = 4$$

? DE QUE FORMA QUEDA AMPLIFICADA?

$$d_1 = 0.15 \quad P_1 = 1 - d_1 = 0.85$$

$$d_2 = 0.85 \quad P_2 = 1 - d_2 = 0.15$$

CON LOS NUEVOS CASOS DE B Y R CALCULO LOS NUEVOS VALORES DE P1 Y P2

$$P_1 = 1 - (1 - (1 - d_1)^R)^B = 0.9778$$

$$P_2 = 1 - (1 - (1 - d_2)^R)^B = 0.0134$$

NOTA

$$\text{FALSOS NEGATIVOS} = 1 - P_1 = 0,022$$

$$\text{FALSOS POSITIVO} = P_2 = 0,0134$$

ES UN LSH BASTANTE PRECISO

(3)

TENGO FALSOS NEGATIVOS

5 AND

3 OR

PARA PODER REDUCIR LOS FALSOS NEGATIVOS

TENEMOS QUE SER MAS LAXOS, AUMENTAR LA CANTIDAD DE ORQ'S.

(4)

QUEDEMOS QUE SI $J(D_1, D_2) = 0,7$

PROBABILIDAD DE QUE SEAN CANDIDATOS (P_1) = 0,9

$J(D_1, D_2) = 0,5$ " " " (P_2) = 0,3

SIENDO $J(D_1, D_2)$ LA SIMILARIDAD DE JACCARD

LA DISTANCIA ES $D_J(D_1, D_2) = 1 - J(D_1, D_2)$

$$D_1 = 0,3$$

$$D_2 = 0,5$$

BUSCAMOS Q Y B PARA FUERZA BRUTA

$$0,9 \geq 1 - (1 - (1 - d_1))^R \quad R = 6$$

$$0,3 \leq 1 - (1 - (1 - d_2))^B \quad B = 19 \quad \left. \begin{array}{l} \\ \end{array} \right\} \text{CUMPLEN}$$

(5)

$$J(D_1, D_2) \geq 0,8$$

$$J(D_1, D_2) \geq 0,9 \quad P_1 \geq 0,95$$

$$D_J(D_1, D_2) = 1 - 0,9 = 0,1 = d_1$$

$$\rightarrow D_J = 1 - J = 0,2 = d_2$$

$$J(D_1, D_2) \leq 0,8 \quad P_2 \leq 0,2$$

CON LA MENOR CANTIDAD DE MH $\rightarrow (B \times B) \rightarrow$ BAJITO

$$P_1 = 1 - (1 - (1 - d_1)^R)^B$$

$$\left. \begin{array}{l} R = 22 \\ B = 29 \end{array} \right\} \text{CUMPLEN}$$

$$P_2 = 1 - (1 - (1 - d_2)^R)^B$$

\downarrow
BUENA MANERA
DE BUSCARLO

(6)

PUNTOS EN EL PLANO = $(2, 3) (3, 4) (24, 30) (21, 32)$

VECTOR AL AZAR = $(5, 3)$

? VALOR W? P_1 Y P_2 SEM

P_3 Y P_4 NO SEM

$$MH_i(x) = \left| \frac{x * v_i + A}{w} \right|$$

\rightarrow DOY VALORES A W HACIENDO
QUE $x * v_i$ PARA P_1 P_2
DEN LO MISMO Y CON P_4
Y P_3 DISTINTOS
 $w = 9,5$

NOTA

9

6 MH

3 CONSTRUCCIONES OR \rightarrow $B = 3$

2 CONSTRUCCIONES AND \rightarrow $R = 2$

6 MH

'USE THE FORCE' = { \$\\$\\$U\$, \$\$US\$, \$U\\$E\$, USE\$_U\$, SELT }

10

VECTORES $X = [1, 3, -1]$

$Z = [2, 4, -1]$

$Y = [-1, -2, -1]$

HIPERPLANOS = $Q_1 = [1, -1, 1]$ $Q_2 = [1, 1, -1]$

$Q_3 = [-1, -1, -1]$ $Q_4 = [1, 1, 1]$

PARA X:

$$[1 \ 3 \ -1] [1 \ -1 \ 1] = -3 = -1$$

$$[1 \ 3 \ -1] [1 \ 1 \ -1] = 5 = 1$$

$$[1 \ 3 \ -1] [-1 \ -1 \ -1] = -3 = -1$$

$$\text{NOT } [1 \ 3 \ -1] [1 \ 1 \ 1] = 3 = 1$$

PABA Y

$$[-1 -2 -1] [1 -1 1] = 0 = 1$$

$$[-1 -2 -1] [1 1 -1] = -2 = -1$$

$$[-1 -2 -1] [-1 -1 -1] = 4 = 1$$

$$[-1 -2 -1] [1 1 1] = -4 = -1$$

PABA Z

$$[2 4 -1] [1 -1 1] = -3 = -1$$

$$[2 4 -1] [1 1 -1] = 7 = 1$$

$$[2 4 -1] [-1, -1, -1] = -5 = -1$$

$$[2 4 -1] [1 1 1] = 5 = 1$$

	x	y	z
R1	-1	1	-1
R2	1	-1	1
R3	-1	1	-1
R4	1	-1	1

$$x - z \rightarrow 4/4$$

$$x - y \rightarrow 0/4$$

$$y - z \rightarrow 0/4$$

11

$$\mathbf{v}_1 = [4, 4, -5, -2, 3] \quad \mathbf{v}_2 = [-3 -2 -4 5 0]$$

$$\mathbf{v}_3 = [3 2 -1 -2 1]$$

HIPÉPLANOS $\mathbf{Q}_1 = [1 1 1 1 -1]$

$$\mathbf{Q}_2 = [-1 1 1 -1 -1]$$

$$\mathbf{Q}_3 = [1 -1 -1 -1 -1]$$

$$\mathbf{Q}_4 = [1 -1 -1 -1 1]$$

$$\mathbf{Q}_5 = [1 -1 -1 -1 1]$$

$$\mathbf{Q}_6 = [-1 1 1 -1 1]$$

$$R = 3 \quad B = 2$$

↓ ↳ 2 TABLAS

3 MH × TABLA

$$[4 4 -5 -2 3] [1 1 1 1 -1] = -2 = -1$$

$$[4 4 -5 -2 3] [-1 1 1 -1 -1] = -6 = -1$$

$$[4 4 -5 -2 3] [1 -1 -1 -1 -1] = 4 = 1$$

$$[4 4 -5 -2 3] [1 -1 -1 -1 1] =$$

$$[4 4 -5 -2 3] [-1 1 1 -1 1] =$$